DOCUMENT RESUME

ED 324 936 FL 018 918

AUTHOR Sedelow, Sally Yeates
TITLE The Interlingual Thesaurus Model for Global Technical Communication: Research Results.
PUB DATE 89
NOTE 12p.; In: Proceedings of the Annual Eastern Michigan University Conference on Languages and Communications for World Business and the Professions (8th, Ann Arbor, MI, March 30-April 1, 1989).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Business Administration Education; *Business Communication; Chinese; Computer Oriented Programs; English; Higher Education; *Intercultural Communication; Language Processing; *Languages for Special Purposes; Machine Translation; Second Language Instruction; *Semantics; *Thesauri; Uncommonly Taught Languages
IDENTIFIERS *Rogets International Thesaurus

ABSTRACT

An interlingual communication support system that shows semantic matches and mismatches among languages being used for any transaction is discussed. The system uses a thesaurus approach to the meaning structure of language, which is hierarchical, but uses a cross-hierarchical approach to access different meanings for selection of the appropriate usage. Thesaurus research has shown a strong correlation between semantic groupings used in "Roget's International Thesaurus" (3rd edition) and the verbal groupings guiding use of the English language, suggesting that the thesaurus may be regarded as the skeleton for English-speaking society's collective associative memory. The approach was tested on the translation of technical information from Chinese to English. A significant problem encountered was the degree of semantic polarity expressed and expected in English in comparison with that of Chinese. The experiment did produce results correlating well with observations made about Chinese in relation to Western languages, and offers a basis for semantic distribution of other languages. The thesaurus approach is seen as having substantial potential for enhancing cross-linguistic comparison. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED324936

THE INTERLINGUAL THESAURAUS MODEL FOR GLOBAL TECHNICAL COMMUNICATION: RESEARCH RESULTS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. M. Vogt

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Dr. Sally Yeates Sedelow
Professor, Computer Science and Adjunct Professor, English
Adjunct Professor, Electronics and Instrumentation
Adjunct Professor, Computer Science
Department of Computer Science
University of Arkansas
Fayetteville, Arkansas 72204

FL018918

THE INTERLINGUAL THESAURUS MODEL FOR GLOBAL TECHNICAL COMMUNICATION: RESEARCH RESULTS

Sally Yeates Sedelow
Professor, Computer Science and Adjunct Professor,
English, University of Arkansas/Little Rock;
Adjunct Professor, Electronics and Instrumentation,
University of Arkansas/Graduate Institute of Technology;
Adjunct Professor, Computer Science,
UA/Fayetteville

The approach taken here toward helping to achieve communicationally adequate languages for world business is that it will often be necessary to proceed in "native tongues," but they should be buttressed by an interlingual communication support system (S. Sedelow, 1987; W. Sedelow, 1989) that shows semantic matches and mismatches among languages being used for any given transaction. The support system envisaged could "back up" machine translation systems or, for speakers/writers with some knowledge of the "other" language(s), could provide most of the assistance needed. Basic to the support system discussed here is a thesaural approach to the meaning structure of language--an approach based upon our extensive investigation of thesaurus modelling and of a particular thesaural instantiation (Roget's International Thesaurus, 3rd Edition (1962)).

To gain an understanding of this approach, a brief description of the type of thesaurus we have studied is in order. First, unlike many thesauri, a Rogetian thesaurus is not set up like a dictionary--that is, it is not an alphabetical listing of words, together with synonyms for each of the words in the listing. Rather, as you probably know, it is organized hierarchically, with the most general categories at the top of a seven or (eight, depending upon

specification of the components) layer 'cake'. For example, an occurrence of the word "perfect" in Roget's International Thesaurus can be shown as follows:

Class Six: Intellect

I. Intellectual Faculties and Properties

L. Conformity to Fact

515. Truth

Adjectives

515.14

faultless, perfect;

As is evident, "perfect" occurs within what we call a Semi-Colon Group, which is the lowest level of the explicit hierarchy in the Thesaurus. At the top of the hierarchy are eight classes, which, taken together, represent one way of grossly dividing many of the words in the English language (all those which appear in the Thesaurus). Such a division, analogous to the classification system (e.g., the Dewey Decimal System, or the Universal Decimal Code, etc.) used for sorting books into shelving areas in a library, is obviously open to debate since we would very likely have variant ways of categorizing the knowledge encoded in English language words. In recognition of this fact, we have used only the lower levels (especially the semi-colon group level) for our research. In addition, our research also has heavily depended upon the implicit structure (in contrast to the explicit hierarchy) of the Thesaurus (W. Sedelow and S. Sedelow, 1987).

The implicit structure is cross-hierarchical and takes advantage of the fact that any given word is likely to appear in more than one place (multi-locality) in the Thesaurus. For example, "perfect" occurs (among other places) in a semi-colon group comprising "complete, perfect, consummate, exhaustive, fully realized," as well as with "faultless" in the example above. An exploration of all the occurrences of "perfect" in the Thesaurus reveals a range of its possible meanings, as made clear by the other words associated with it. One could use the textual context of a potentially ambiguous word in order to "focus in" or "zoom in" on the appropriate meaning, as defined by a given location in the Thesaurus. It was with such a possibility in mind that our student, Professor Robert Bryan (1973), developed a formal, mathematical model of thesauri, using the third edition of Roget's International Thesaurus for the instantiation. Part of the model consists of the definition of "chains," ranging from Type 1 through Type 10, which provide formally defined rules (which therefore can be followed by a computer program) for moving around within the implicit structure of a thesaurus. Not all the chains depend upon multi-locality; in fact, the Type 1 chain is so general (any collection of entries in the Thesaurus) as to be of no evident utility for language understanding. At the other extreme, the Type 10 chain demands a "strong" connection between links in the chain, so that two categories must be linked together by at least two words occurring in both categories, and the words, in turn, are linked together by at least the two categories. Bryan also mathematically defines the Star (chains radiating out from a given entry) and the Neighborhood, which consists of whatever the arms of the Star cover.

Given both components--explicit and implicit--of a topological model of thesauri, we were encouraged to put the instantiation into fully computer-accessible form, so as to conduct research in a rigorous, automated way. To do so required editing so as, for example, to make such properties as elisions (for instance, "jump up or off") explicit; some of this editing was in fact done by computer, wherever modifications could be rulefully (formally) described. For those who have an interest in the necessary preparation of such databases, a reasonably full description of our effort appears in "The Lexicon in the Background" (S. Sedelow and W. Sedelow, 1986) and full documentation is to be found in Harris (1973) and Taylor (1973).

Our research on the Thesaurus has ranged over a number of different application areas; that research has demonstrated the strong correlation between semantic groupings (semantic spaces or sub-spaces) in the Thesaurus and those verbal groupings which guide our use of the English language (S. Sedelow, 1985, ASIS; W. Sedelow, 1985, ASIS; S. Sedelow and W. Sedelow, 1985, 1989). In fact, our empirical research suggests that the Thesaurus might be accurately regarded as the skeleton for English-speaking society's collective associative memory. For this conference, a highly relevant project is one inspired by an inquiry from a U.S. business executive involved in efforts to transfer technology between primarily Chinese speaking and primarily English speaking engineers. The business executive observed that although communication was thought by both groups to have been achieved in any given session, in fact it was subsequently clear that affirmative headshaking was premature and misleading. It occurred to us that there ought to be a communication support system that could display--

perhaps on a large screen--either actual mismatches between the languages or potential problems posed by potentially ambiguous terms (many-to-one and one-to-many mappings). Perhaps, we thought, a display could exhibit the actual convergence, within a semi-colon group or other thesaural semantic sub-space, of a number of contextual terms, and then that displayed convergence would help to reveal to the non-native speaker with a shaky knowledge of English where a given (ambiguous) term in the discourse in question might be placed semantically. At the least, it would be useful at the outset to be able to show non-matches or mismatches between one language and the other, first at the global level and then narrowed down to the specific application domain (semantic infrastructure).

Our first approach to this problem was to use Karlgren's (1974 reprint) simplicia (which was meant to represent "conceptual roots" for Chinese ideographs) as a means of comparing Chinese and English. To make the comparison, we distributed the simplicia "over" the Thesaurus in an effort to determine whether there were sections of the Thesaurus which strongly intersected with Karlgren's categories, as well as sections which correlated either weakly or not at all. When looking at the results, of course it was necessary to bear in mind the disparate sizes of the Thesaurus (210,000) entries and of the simplicia (6159). Relatively small size notwithstanding, the simplicia according to Karlgren were, taken as a whole, more wide-ranging than the stem words included in the largest existing English-Chinese dictionaries at the time Karlgren was compiling his dictionary. Included with each stem word in the Karlgren list is a group of English words which provided at least some semantic equivalence with the meaning of the Chinese stem word.

To make the comparison we took the English words approximating each stem word and looked for matches in the Thesaurus in order to see what the pattern of "hits" would be. Even at the top level there are significant differences in the percentage of conceptual matches between the two languages. Abstract Relations had the highest percentage of matches, 17.52%, and Physics, the lowest, 11.23% (N.B. that all the percentages will be low, given the disparity in size between the simplicia and the Thesaurus). Abstract Relations owed its #1 position to two subcategories, Substantiality and State, which had very high (38%) "hit" rates. These subcategories contain very general terms such as body, mass, solidity, something, thing, and so on--terms which are abstract and not, surprisingly, represented in both languages. We did notice, though, when examining Substantiality, that its opposite, Unsubstantiality, had a low percentage of "hits" (11.95%). We then looked at a number of other opposites and saw the same pattern: Relation with a 17.05% match rate vs. Irrelation with 8.00%; Similarity with 18.75% vs. Dissimilarity with 9.23%; Agreement with 22.76% vs. Disagreement with 5.95%; and Superiority with 16.61% vs. Inferiority with 11.21%. These data lend support to the observation that many Indo-European languages, including English, place more emphasis upon semantic polarities than do Eastern languages, including Chinese.

At the gross level, differential emphases upon polarities represent the kind of information that should be made available in advance to speakers trying to communicate across linguistic boundaries. Clearly, for example, strong expectations for "yes" or "no" (binary) responses are set up by a general-language orientation toward polarities; whereas, a user of a language without such an emphasis is much more likely to think of points or intervals on a

continuum (multi-valued 'logic'). Such a disparity in language use and expectation as to language use holds a rich promise for confusion in business transactions--one side may well suppose that a non-negative answer means "yes," while the answerer may well neither mean "no," nor "yes," but something on a continuum in between, which may be exactly what the answerer said. The hearer, thinking, he heard "yes," may then be disappointed and accuse the answerer of bad faith, which will outrage the answerer, who spoke accurately and in good faith--but (relative to) a different language.

There is not space here to go into detailed omparisons. But it would be clear that were one seeking to communicate (at the time of Karlgren's compilation of the simplicia) ideas relating to Rockets and Flying Missiles, a 3.74% match rate would suggest that much supplementation would be needed, and even more so for the .46% match rate for Astronautics. Terms relating to Light had a 14.76% rate, but Electricity and Electronics only 4.96%. In these instances, one would then look further to see whether there was simply a paucity of terms on the one side relative to the other, or whether both sides had many terms which simply didn't match at the level of semantic infrastructure. It would be important to be aware of both situations (i.e., of gaps and of mismatches) so as to know how the communication problem could best be addressed.

This experiment, with which we had the assistance of two excellent Chinese students, did produce results which correlate well with observations made about the Chinese language vis-a-vis Western languages (Bloom, 1981). And given the topological model of the Thesaurus, we have a structure of English semantic space against which to distribute other languages (S. Sedelow and Mooney, 1988;

W. Sedelow, 1988, ASIS). Thus, we have a very strong foothold on the no doubt still rocky path twoard cross-linguistic comparisons utilizable in the Interlingual Communication Support System (ICSS) environment we have in mind.

The prospects for the long-term utility of this approach seem particularly rich, given the associative structure of the thesaural model. Recent research in the neurosciences (Sejnowski, et al., 1988), as well as developments in the areas of "genetic computing" (Holland, 1975, Walbridge, 1989), and "natural computing" (Richards, 1988), would seem to suggest a compatability between an associative structure and human neural function. Further, we believe the associative structure very likely makes accessible data now being labelled with such terms as "belief systems," "presupposition," "frames," "stereotypes," etc. --the kinds of information, in short, which a businessman carries along as background when he or she launches into negotiations with counterparts from a different language community.

In addition to a more general kind of background information (perhaps characterizable as cultural), the Thesaurus as it stands contains far more domain-targeted information than we had at first expected. For example, we have tested it against a collection of medical abstracts to see whether it would "retrieve" the structuring concepts in the abstracts, and it performed well in this exercise (Brady, 1988). Future plans for research on the Thesaurus include an effort to add automatically domain-specific terms to appropriate semantic subspaces in the Thesaurus. If successful, such augmentation would provide both a domain-transcendent and domain-specific language resource against which could be modelled the relevant 'other' language(s) in any given commercial exchange.

In conclusion, an obvious point should be stressed--this large, associative semantic net of English is already available and already extensively studied and tested. Thus, it has both short-term and long-term potential for the tasks at hand.

Bibliography

Bloom, Alfred H. 1981. Linguistic Shaping of Thought: A Study in The Impact of Language on Thinking in China and the West. New Jersey: Laurence Erlbaum.

Brady, John. 1988. "ICSS (Interlingual Communication Support System) and a Wittgensteinian Language Game," Proceedings, European Studies Conference, University of Nebraska/Omaha (In Press).

Bryan, Robert. 1973. "Abstract Thesauri and Graph Theory Applications in Thesaurus Research," in S. Sedelow, et al., Automated Language Analysis 1972-1973, pp. 45-89. University of Kansas Departments of Computer Science and Linguistics.

Harris, Herbert R. 1973. "The Conversion of Roget's International Thesaurus to an Automated Data Base," in S. Sedelow, et al., Automated Language Analysis 1972-1973. University of Kansas Departments of Computer Science and Linguistics.

Holland, John H. 1975. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. Ann Arbor: University of Michigan.

Karlgren, Bernhard. 1974. Analytical Dictionary of Chinese and Sino-Japanese, Dover.

Richards, Whitman, ed. 1988. Natural Computation. Cambridge, Mass., MIT Press.

Roget's International Thesaurus. 1963, Third Edition. New York: Thomas Y. Crowell.

Sedelow, Sally Yeates. 1985. "Computational Literary Thematic Analysis: The Possibility of a General Solution," in Carol Parkhurst, ed., Proceedings, 48th Annual Meeting of the American Society for Information Science, 22: 359-362.

Sedelow, Sally Yeates. 1987. "An Interlingual Communication Support System (ICSS) Example Re Chinese/English Classroom Instruction," Proceedings, Methods III, International Conference on Foreign Language Teaching; University of Northern Iowa, Cedar Falls (In Press).

Sedelow, Sally Yeates, with Donna Weir Mooney, 1988. "Knowledge Retrieval from Domain-Transcendent Expert Systems: II. Research Results," Proceedings, 51st Annual Meeting of the American Society for Information Science, 25: 209-212.

Sedelow, Sally Yeates and Walter A. Sedelow, Jr. 1988. "The Application to Business/Industry of the Domain-Transcendent Expert System Building Methodology," in R. Vedder, et al., eds., Proceedings, Conference on the Impact of Artificial Intelligence on Business and Industry; Information Systems Research Center, University of North Texas, Denton, pp. 98-105.

Sedelow, Sally Yeates and Walter A. Sedelow, Jr. 1989. "Artificial Intelligence, Expert Systems, and Productivity," in Whitney and Ochsman, eds., Psychology and Productivity: Bringing Together Research and Practice. Plenum Press, pp. 51-68.

Sedelow, Walter A. 1987. "The Interlingual Communication Support System (ICSS): Underlying Concepts and Procedures," Proceedings, Methods III, International Conference on Foreign Language Teaching, University of Northern Iowa, Cedar Falls (In Press).

Sedelow, Walter A. 1988. "Knowledge Retrieval from Domain-Transcendent Expert Systems: I. Some Concepts from Cognitive Robotics," Proceedings, 48th Annual Meeting of the American Society for Information Science, 22: 363-366.

Sedelow, Walter A., Jr. and Sally Yeates Sedelow. 1987. "Semantic Space," in Winfred Lehmann, ed., Computers and Translation, 2: 231-242.

Sejnowski, Terrence J., Christof Koch, and Patricia S. Churchland. 1988. "Computational Neuroscience," Science, 241 (9 September 1988), pp. 1299-1306.

Taylor, Scott. 1973. "Handling of Bracketed Information," in S. Y. Sedelow, et al., Automated Language Analysis, University of Kansas Departments of Computer Science and Linguistics.

Walbridge, Charles. 1989. "Genetic Algorithms: What Computers Can Learn from Darwin,": Technology Review, M.I.T., January, pp. 47-53.

END

U.S. Dept. of Education

Office of Education
Research and
Improvement (OERI)

ERIC

Date Filmed

March 29, 1991